



中华人民共和国国家标准化指导性技术文件

GB/Z 243—2026

人工智能 端侧大模型推理引擎技术要求

Artificial intelligence—Technical requirements for on-device large-scale
model inference engine

2026-08-27 发布

国家市场监督管理总局 发布
国家标准化管理委员会

目 次

前言 III

引言 IV

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 1

5 参考架构 2

6 技术要求 2

 6.1 数据 2

 6.2 轻量化部署 4

 6.3 异构软硬件适配 6

 6.4 推理计算加速 6

 6.5 推理内存优化 8

参考文献 10

前 言

本文件为规范类指导性技术文件。

本文件按照 GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。文件的发布机构不承担识别专利的责任。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：上海交通大学、中国电子技术标准化研究院、浙江大学、淘宝(中国)软件有限公司、联想(北京)有限公司、中兴通讯股份有限公司、中国电信集团有限公司、青岛国创智能家电研究院有限公司、济南浪潮数据技术有限公司、中国电力科学研究院有限公司、中国移动通信集团终端有限公司、中国电信股份有限公司上海分公司、蔚来控股有限公司、上海壁仞科技股份有限公司、OPPO 广东移动通信有限公司、天津大学、南京大学。

本文件主要起草人：吴帆、牛超越、吴飞、鲍薇、吕承飞、吴杰、张东、王晔、王茜莺、周飞、郝建业、许辉阳、龚勃、刘彪、卢龙、丁瑞全、秦征、陈贵海、张慷、朱文印、董建、徐洋、崔卓、黄正翔、郑臻哲、杨易、张圣宇、郑若琳、刘晓璇、王召德、王晓辉、邢冯、陈晓春、朱姝、刘生钟、柳宁、陈静远、范鹤鹤、陈志文、康峰、李沁雅、章敏、魏颖、周煊赫、朱霖潮、贺可强、刘泽民。

引 言

端侧大模型推理引擎作为人工智能技术演进中的关键方向,致力于将高性能的大语言模型和多模态大模型部署于移动端设备,以满足对低延迟响应、离线处理、低成本、隐私保护等日益增长的实际需求。相较于传统依赖云侧大模型推理部署方式,端侧推理引擎能在保持智能能力的同时,显著提升系统的自主性、实时性与鲁棒性,已成为产业界与学术界高度关注的研究热点,并在智能终端、物联网、自动驾驶、工业控制等领域展现出广阔的应用前景与实际价值。本文件旨在规范当前端侧大模型推理引擎的系统架构与关键技术路径,推动端侧资源约束下的模型压缩、高效异构计算与调度优化等能力的标准化建设,提升端侧智能系统的普适部署能力和运行效率,进一步拓展大模型技术在终端侧的可持续发展与规模化落地,助力智能生态体系的全面升级与广泛赋能。

人工智能 端侧大模型推理引擎技术要求

1 范围

本文件确立了功能视角下的端侧大模型推理引擎参考架构,规定了大模型在端侧部署的技术要求。本文件适用于端侧大模型推理引擎的设计、开发、测试和部署。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 45288.1—2025 人工智能 大模型 第1部分:通用要求

3 术语和定义

GB/T 41867—2022、GB/T 45288.1—2025 界定的以及下列术语和定义适用于本文件。

3.1

端侧大模型推理引擎 on-device large-scale model inference engine

部署在端侧设备上,用于对已训练的大模型进行加载、调度和执行,利用设备本地计算资源实现高效推理,并支撑特定端侧智能任务运行的基础软件。

3.2

模型剪枝 model pruning

通过删除冗余参数或结构,降低模型规模与计算成本的模型轻量化方法。

3.3

知识蒸馏 knowledge distillation

用复杂大模型(教师模型)指导轻量化模型(学生模型),将核心知识压缩迁移,在降低模型资源占用的同时,尽可能保留推理精度的模型轻量化方法。

4 缩略语

下列缩略语适用于本文件。

CPU:中央处理器(Central Processing Unit)

GPU:图形处理器(Graphic Processing Unit)

NPU:神经处理器(Neural Processing Unit)

PTQ:后训练量化(Post-Training Quantization)

QAT:量化感知训练(Quantization Aware Training)